PCT/JP03/09678

日 本 国 特 許 庁
JAPAN PATENT OFFICE

30.07.03

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年 ７月３０日

出願番号
Application Number: 特願２００２－２２１６９８

[ST.10/C]: [ＪＰ２００２－２２１６９８]

出願人
Applicant(s): ソニー株式会社

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN COMPLIANCE WITH RULE 17.1 (a) OR (b)

２００３年 ６月 ９日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号 出証特２００３－３０４４５２０

【書類名】　特許願
【整理番号】　0290377102
【提出日】　平成14年 7月30日
【あて先】　特許庁長官殿
【国際特許分類】　H04N　5/44
【発明者】
　【住所又は居所】　東京都品川区北品川６丁目７番３５号　ソニー株式会社内
　【氏名】　木村　仁史
【発明者】
　【住所又は居所】　東京都品川区北品川６丁目７番３５号　ソニー株式会社内
　【氏名】　大沼　顕介
【発明者】
　【住所又は居所】　東京都品川区北品川６丁目７番３５号　ソニー株式会社内
　【氏名】　市岡　秀俊
【特許出願人】
　【識別番号】　000002185
　【氏名又は名称】　ソニー株式会社
【代理人】
　【識別番号】　100122884
　【弁理士】
　【氏名又は名称】　角田　芳末
　【電話番号】　03-3343-5821
【選任した代理人】
　【識別番号】　100113516
　【弁理士】
　【氏名又は名称】　磯山　弘信

【電話番号】　03-3343-5821

【手数料の表示】

【予納台帳番号】　176420

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　0206460

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　キーワードの自動抽出装置及び方法

【特許請求の範囲】

【請求項１】　コンテンツのタイトル文字列情報から、サブジャンルを示す文字列が登録されている第１のキーワード辞書を用いてキーワードの抽出を行う第１の抽出手段と、

前記コンテンツの詳細文字列情報から、人名が登録されている第２のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第２の抽出手段と

を備えたことを特徴とするキーワード自動抽出装置。

【請求項２】　請求項１に記載のキーワード自動抽出装置において、

前記第１の抽出手段は、第１のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、所定の除外文字列辞書に登録されている文字列を除外した部分からキーワードを抽出することを特徴とするキーワード自動抽出装置。

【請求項３】　請求項１に記載のキーワード自動抽出装置において、

前記第１の抽出手段は、第１のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、平仮名，カタカナ，漢字，数字，アルファベット以外の特殊文字で区切られている文字列をキーワードとして抽出することを特徴とするキーワード自動抽出装置。

【請求項４】　請求項１に記載のキーワード自動抽出装置において、

前記第２の抽出手段は、前記第２のキーワード辞書を用いてキーワードを抽出した前記詳細文字列情報の残りの部分のうち、所定の除外文字列辞書に登録されている文字列を除外した部分から、字種切り法を利用したキーワードの抽出を行うことを特徴とするキーワード自動抽出装置。

【請求項５】　請求項１に記載のキーワード自動抽出装置において、

前記第２の抽出手段は、字種切り法を利用しつつ、カタカナとアルファベットとを同一字種として扱うとともに、‘・’（中点）は、その直前の文字がカタカナ，アルファベットである場合にはそれぞれカタカナ，アルファベットとして扱うことを特徴とするキーワード自動抽出装置。

【請求項６】　請求項１に記載のキーワード自動抽出装置において、

前記第２のキーワード辞書をネットワーク経由でダウンロードする手段をさらに備え、前記第２の抽出手段は前記ダウンロードされた第２のキーワード辞書を用いることを特徴とするキーワード自動抽出装置。

【請求項７】　コンテンツのタイトル文字列情報から、サブジャンルを示す文字列が登録されている第１のキーワード辞書を用いてキーワードの抽出を行う第１のステップと、

前記コンテンツの詳細文字列情報から、人名が登録されている第２のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第２のステップと

を有することを特徴とするキーワード自動抽出方法。

【請求項８】　請求項７に記載のキーワード自動抽出方法において、

前記第１のステップで、第１のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、所定の除外文字列辞書に登録されている文字列を除外した部分からキーワードを抽出することを特徴とするキーワード自動抽出方法。

【請求項９】　請求項７に記載のキーワード自動抽出方法において、

前記第１のステップで、第１のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、平仮名，カタカナ，漢字，数字，アルファベット以外の特殊文字で区切られている文字列をキーワードとして抽出することを特徴とするキーワード自動抽出方法。

【請求項１０】　請求項７に記載のキーワード自動抽出方法において、

前記第２のステップで、前記第２のキーワード辞書を用いてキーワードを抽出した前記詳細文字列情報の残りの部分のうち、所定の除外文字列辞書に登録されている文字列を除外した部分から、字種切り法を利用したキーワードの抽出を行うことを特徴とするキーワード自動抽出方法。

【請求項１１】　請求項７に記載のキーワード自動抽出方法において、

前記第２のステップで、字種切り法を利用しつつ、カタカナとアルファベットとを同一字種として扱うとともに、‘・’（中点）は、その直前の文字がカタカナ，アルファベットである場合にはそれぞれカタカナ，アルファベットとして扱

うことを特徴とするキーワード自動抽出方法。

【請求項１２】　請求項７に記載のキーワード自動抽出方法において、

前記第２のキーワード辞書をネットワーク経由でダウンロードするステップをさらに有し、前記第２のステップでは前記ダウンロードした第２のキーワード辞書を用いることを特徴とするキーワード自動抽出方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、ＥＰＧ（Electronic Program Guide：電子番組ガイド）情報のようなコンテンツのタイトル文字列情報及び詳細文字列情報から、キーワードを自動的に抽出する装置及び方法に関する。

【０００２】

【従来の技術】

近年本格化しているディジタルテレビジョン放送では、番組の映像・音声データとともに、番組のタイトルを示す情報（タイトル文字列情報）や番組の詳細を説明する情報（詳細文字列情報）や番組のジャンルを示す情報等を含んだＥＰＧ情報が、放送局から送信される。ディジタル放送に対応したテレビジョン受信機では、このＥＰＧ情報に基づいて画面上に電子番組ガイドを表示させることができる。

【０００３】

また、アナログテレビジョン放送でも、こうしたＥＰＧ情報が送信されているものがある。

【０００４】

ユーザーは、見たい番組を検索する場合、この電子番組ガイドを利用して、大まかなジャンル（例えばスポーツ，ドラマ等）を選んだ後、タイトルから検索したり、詳細文字列情報を読んで検索したりしている。

【０００５】

しかし、番組のタイトルの付け方は千差万別であるので、タイトルからの検索はユーザーにとって必ずしも行いやすいものではない。また、番組の詳細文字列

情報は文章の形式で記述されており何ページにも亘ることが少なくないので、詳細文字列情報からの検索もユーザーにとって面倒である。

【０００６】

これに対し、例えば芸能人名等のキーワードを用いて番組を検索できるようにすれば、ユーザーにとって検索が非常に容易になる。しかるに、現在放送局から送信されるＥＰＧ情報には、キーワードは独立して含まれていない。そのため、キーワードを用いた検索を可能にするためには、ＥＰＧ情報からキーワードを抽出することが必要になる。

【０００７】

従来、このキーワードの抽出方法としては、テレビジョン受信機に表示された電子番組ガイド中の詳細文字列情報の文章の中から、ユーザーが、キーワードとして決定したい文字列の先頭及び末尾の語をカーソル等で指定するという方法が存在していた。

【０００８】

【発明が解決しようとする課題】

しかし、この従来の抽出方法では、ユーザーが自らキーワード指定のための操作を行わなければならないので、煩雑であるとともに、多数のキーワードを短時間に抽出することは困難である。

【０００９】

他方、一般的なキーワードの自動抽出方法としては、日本語形態要素解析という方法が知られている。しかし、この方法は、プログラムのサイズや使用する辞書のサイズが非常に大きいとともに、ＣＰＵに大きな負荷をかけてしまう。したがって、テレビジョン受信機のようにＣＰＵの処理能力やメモリの容量がさほど大きくない家電製品でこの方法を用いることは、極めて非効率的である。

【００１０】

さらに、一般的なキーワードの自動抽出方法としては、字種切り法という方法も知られている。この方法は、漢字・カタカナ・平仮名・アルファベット・数字等の字種の違いを検出することによってキーワードを抽出するものである。しかし、この字種切り法だけでは、番組を検索するためのキーワードの抽出を精度よ

く行うことはできない。すなわち、苗字が漢字で名前が平仮名やカタカナの芸能人名（例えば‘石田あかり’というような名称）は、苗字と名前とが分割されてしまうので抽出することができない。また、名前がアルファベットで表記され苗字がカタカナで表記された外国人名や名前と苗字との間に‘・’（中点）が挿入された外国人名（例えば‘Ｂ・ドゥーリー’というような名称）も、名前と苗字とが分割されてしまうので抽出することができない。

【００１１】

本発明は、上述の点に鑑み、ＣＰＵの処理能力やメモリの容量がさほど大きくない家電製品でも、ＥＰＧ情報のようなコンテンツのタイトル文字列情報及び詳細文字列情報から、ユーザーがコンテンツを検索するためのキーワードを、効率よく且つ精度よく自動的に抽出できるようにすることを課題としてなされたものである。

【００１２】

【課題を解決するための手段】

この課題を解決するために、本出願人は、コンテンツのタイトル文字列情報から、サブジャンルを示す文字列が登録されている第１のキーワード辞書を用いてキーワードの抽出を行う第１の抽出手段と、このコンテンツの詳細文字列情報から、人名が登録されている第２のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第２の抽出手段とを備えたキーワード自動抽出装置を提案する。

【００１３】

このキーワード自動抽出装置では、コンテンツのタイトル文字列情報（例えばデジタルテレビジョン放送ではＥＰＧ情報中のタイトル文字列情報）からは、サブジャンルを示す文字列が登録されている第１のキーワード辞書を用いてキーワードが抽出される。

【００１４】

また、コンテンツの詳細文字列情報（例えばデジタルテレビジョン放送ではＥＰＧ情報中の詳細文字列情報）からは、人名が登録されている第２のキーワード辞書を用いてキーワードが抽出されるとともに、字種切り法を利用したキーワー

ドの抽出も行われる。その際、第２のキーワード辞書に登録されている人名であれば、苗字が漢字で名前が平仮名やカタカナの人名もキーワードとして抽出される。また、第２のキーワード辞書に登録されていない人名でも、字種切り法を利用することによってキーワードとして抽出される。

【００１５】

このように、タイトル文字列情報からのキーワードの抽出と詳細文字列情報からのキーワードの抽出とを、それぞれの情報に合せて互いに異なるキーワード辞書とルール（字種切り法を利用するか否か等）で行うことにより、小さなサイズのプログラムや辞書で精度よくキーワードを抽出することができる。

【００１６】

これにより、ＣＰＵの処理能力やメモリの容量がさほど大きくない家電製品でも、ＥＰＧ情報のようなコンテンツのタイトル文字列情報及び詳細文字列情報から、ユーザーがコンテンツを検索するためのキーワードを、効率よく且つ精度よく自動的に抽出することができるようになる。

【００１７】

なお、このキーワード自動抽出装置において、一例として、第１の抽出手段は、第１のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、所定の除外文字列辞書に登録されている文字列を除外した部分からキーワードを抽出することが好適である。

【００１８】

それにより、タイトルに含まれていることのある文字列のうち、コンテンツを検索するためには不適切な（一般的過ぎるような）文字列がキーワードに含まれることを防止することができる。したがって、ユーザーは、抽出されたキーワードを用いて、コンテンツを一層効率的に検索することができるようになる。

【００１９】

さらに、このキーワード自動抽出装置において、一例として、第１の抽出手段は、第１のキーワード辞書に登録されている文字列を含むタイトル文字列のうち、平仮名，カタカナ，漢字，数字，アルファベット以外の特殊文字で区切られている文字列をキーワードとして抽出することが好適である。

【００２０】

それにより、こうした特殊文字で区切られていないタイトルについては、そのタイトルに含まれる複数の文字列がばらばらのキーワードとして抽出されることなく、そのタイトルそのものがそのままの形でキーワードとして抽出されるようになる。

【００２１】

こうした特殊文字で区切られていないタイトルは、そのタイトルに含まれる個々の文字列は意味が広すぎてコンテンツ検索のためのキーワードとしてあまり役立たず（検索結果が非常に多くなり）、タイトルそのものとしてはじめてコンテンツの効率的な検索のためのキーワードとして役立つことが多い。したがって、ユーザーは、抽出されたキーワード（タイトルそのもの）を用いて、コンテンツを一層効率的に検索することができるようになる。

【００２２】

また一方では、特殊文字で区切られているタイトルについては、特殊文字で区切られている個々の文字列がそれぞれキーワードとして抽出されるようになる。

【００２３】

特殊文字（例えばスペースや‘×’等）で区切られているタイトルは、その特殊文字で区切られている個々の文字列がそれぞれコンテンツ検索のためのキーワードとして役立ち、タイトルそのものとしては限定されすぎてコンテンツ検索のためのキーワードとしてあまり役立たない（検索結果がゼロまたは非常に少なくなる）ことが多い。したがって、ユーザーは、抽出されたキーワード（特殊文字で区切られている個々の文字列）を用いて、やはりコンテンツを一層効率的に検索することができるようになる。

【００２４】

さらに、このキーワード自動抽出装置において、一例として、第２の抽出手段は、第２のキーワード辞書を用いてキーワードを抽出した詳細文字列情報の残りの部分のうち、所定の除外文字列辞書に登録されている文字列を除外した部分から、字種切り法を利用したキーワードの抽出を行うことが好適である。

【００２５】

それにより、詳細文字列情報に含まれていることのある文字列のうち、コンテンツを検索するためには不適切な文字列がキーワードに含まれることを防止することができる。したがって、ユーザーは、抽出されたキーワードを用いて、コンテンツを一層効率的に検索することができるようになる。

【００２６】

さらに、このキーワード自動抽出装置において、一例として、第２の抽出手段は、字種切り法を利用しつつ、カタカナとアルファベットとを同一字種として扱うとともに、‘・’（中点）は、その直前の文字がカタカナ，アルファベットである場合にはそれぞれカタカナ，アルファベットとして扱うことが好適である。

【００２７】

それにより、名前がアルファベットで表記され苗字がカタカナで表記された外国人名や名前と苗字との間に‘・’（中点）が挿入された外国人名も、キーワードとして抽出することができるようになる。

【００２８】

さらに、このキーワード自動抽出装置において、第２のキーワード辞書をネットワーク経由でダウンロードする手段をさらに備え、第２の抽出手段はこのダウンロードされた第２のキーワード辞書を用いることが好適である。

【００２９】

それにより、第２のキーワード辞書として、最新の辞書（最近有名になったばかりの人の名称も登録されている辞書）を用いてキーワードを抽出することができるようになる。

【００３０】

次に、本出願人は、コンテンツのタイトル文字列情報から、サブジャンルを示す文字列が登録されている第１のキーワード辞書を用いてキーワードの抽出を行う第１のステップと、このコンテンツの詳細文字列情報から、人名が登録されている第２のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第２のステップとを有するキーワード自動抽出方法を提案する。

【００３１】

このキーワード自動抽出方法によれば、前述の本発明に係るキーワード自動抽出装置について説明したのと全く同様にして、ＣＰＵの処理能力やメモリの容量がさほど大きくない家電製品でも、ＥＰＧ情報のようなコンテンツのタイトル文字列情報及び詳細文字列情報から、ユーザーがコンテンツを検索するためのキーワードを、効率よく且つ精度よく自動的に抽出することができるようになる。

【００３２】

【発明の実施の形態】

以下、デジタルテレビジョン放送の番組を記録・再生する装置に本発明を適用した例について、図面を用いて説明する。

【００３３】

図１は、本発明を適用した番組記録再生装置を含むデジタルテレビジョン放送受信システムの概要を示す図である。テレビジョン放送局から送信されたデジタル放送信号が、アンテナ１で受信されて番組記録再生装置２に入力する。番組記録再生装置２は、ディスプレイ及びスピーカを含む表示装置３に接続されるとともに、インターネット４に接続されている。

【００３４】

図２は、番組記録再生装置２のハードウェア構成を示すブロック図である。この番組記録再生装置２では、チューナ１１，復調器１２，デスクランブラ１３，多重分離器１４が順に接続されるとともに、多重分離器１４に対して映像デコーダ１５，映像信号処理回路１７と音声デコーダ１６，Ｄ／Ａ変換器１８とがそれぞれ順に接続されている。

【００３５】

また、チューナ１１～Ｄ／Ａ変換器１８，ＣＰＵ１９，ＲＯＭ２０，メインメモリ（ＲＡＭ）２１，フラッシュメモリ２２，リモートコントローラ用のインターフェース２３，ＨＤＤ（ハードディスクドライブ）用のインターフェース２４，インターネット接続用の通信インターフェース２５が、互いにシステムバス２６で結ばれている。インターフェース２４には、テレビ番組を録画するためのＨＤＤ（ハードディスクドライブ）２７が接続されている。

【００３６】

この番組記録再生装置２に付属したリモートコントローラ（以下リモコンと呼ぶ）２８には、通常のデジタル放送用テレビジョン受信機に付属したリモコンにおけるのと同じ各種の操作釦（電源釦や、選局釦や、録画予約釦や、再生釦や、ＥＰＧ画面上で選択を行うための方向キーや決定キー等）が設けられている。

【００３７】

テレビ番組の視聴時には、番組記録再生装置２に入力したデジタル放送信号は、リモコン２８の選局操作に基づいてチューナ１１で周波数帯を選択された後、復調器１２で復調され、デスクランブラ１３でスクランブルを解かれた後、多重分離器１４で、複数チャンネル分の番組の映像・音声データのパケットやＥＰＧ情報のパケットに分離される。

【００３８】

この複数チャンネル分のテレビ番組の映像・音声のパケットのうち、リモコン２８の選局操作に基づいて抽出した１チャンネル分のパケットの映像，音声データが，それぞれ映像デコーダ１５，音声デコーダ１６によりＭＰＥＧ－２Ｖｉｄｅｏ，ＭＰＥＧ－２Ａｕｄｉｏで復号される。また、ＥＰＧ情報のパケットはＣＰＵ１９に送られる。

【００３９】

そして、映像デコーダ１５で復号された映像信号や、ＥＰＧ情報を用いてＣＰＵ１９で作成された電子番組ガイド表示用の映像信号が、映像信号処理回路１７でＮＴＳＣ方式への変換やミキシング等を施され、映像出力端子２９から出力して図１の表示装置３に送られる。

【００４０】

また、音声デコーダ１６で復号された音声信号が、Ｄ／Ａ変換器１８でアナログ変換され、音声出力端子３０から出力して図１の表示装置３に送られる。

【００４１】

ＣＰＵ１９は、ＲＯＭ２０に格納されたプログラムやデータに基づき、メインメモリ２１をワーキングメモリとして用いてこの番組記録再生装置２全体を制御する。

【００４２】

ＣＰＵ１９が行う処理には、リモコン２８の選局操作に基づくテレビ番組の視聴時の処理や、リモコン２８の録画予約操作に基づくＨＤＤ２７へのテレビ番組の録画処理の他に、キーワードの自動抽出処理がある。

【００４３】

ＲＯＭ２０には、このキーワードの自動抽出処理で用いるための辞書として、タイトル用キーワード辞書，タイトル用除外文字列辞書，詳細情報用キーワード辞書及び詳細情報用除外文字列辞書が格納されている。

【００４４】

タイトル用キーワード辞書には、‘プロ野球’，‘ゴルフ’，‘サッカー’，‘温泉’，‘囲碁’，‘将棋’，‘映画’等のサブジャンル（ＥＰＧ情報中のジャンル情報による‘スポーツ’といったような大まかなジャンルよりも細かいジャンル）を示す文字列や、‘恋’，‘愛’といった文字列や、プロ野球の球団名の文字列のように、番組のタイトルに含まれていることの多い文字列のうち、番組を検索するために有効且つ重要な文字列が登録されている。

【００４５】

タイトル用除外文字列辞書には、‘映画’，‘ＢＳ’，番組表独特の記号（例えば、ニュース番組を表す、Ｎを四角の枠で囲んだ記号）といったように、番組のタイトルに含まれていることのある文字列のうち、番組を検索するためのキーワードとしては一般的過ぎる文字列が登録されている。

【００４６】

詳細情報用キーワード辞書には、テレビジョン番組に登場することの多い有名人（芸能人，スポーツ選手，政治家，文化人等）の名称のうち、平仮名のみ，平仮名と漢字との組み合わせ，平仮名とカタカナとの組み合わせ，漢字とカタカナとの組み合わせ，２文字以下の漢字のみ，６文字以上の漢字のみの名称の文字列がそれぞれ登録されている。また、詳細情報用キーワード辞書には、例えば‘温泉’というような、ＥＰＧ情報中の詳細文字列情報に含まれていることの多い文字列のうち、人名以外の文字列であって番組を検索するためのキーワードとして適切な文字列も登録されている。

【００４７】

詳細情報用除外文字列辞書には、‘ゲスト’，‘以上’，‘監督’のように、ＥＰＧ情報中の詳細文字列情報に含まれていることの多い文字列のうち、番組を検索するためのキーワードとしては不適切な文字列が登録されている。

【００４８】

なお、ＣＰＵ１９は、詳細情報用キーワード辞書に関しては、インターネット経由で専用のサイトから最新のもの（最近有名になったばかりの人の名称等が登録されているもの）をダウンロードしてフラッシュメモリ２２にも記憶させる。

【００４９】

また、ＣＰＵ１９は、キーワードの自動抽出処理を行う前提として、ユーザーの選局操作時やユーザーの録画予約操作に基づく録画時に多重分離器１４から送られたＥＰＧ情報のパケットをフラッシュメモリ２２に記憶させる。

【００５０】

図３，図４は、ＣＰＵ１９が実行するキーワードの自動抽出処理を示すフローチャートである。このうち、図３は、タイトル文字列情報からキーワードを抽出する処理であり、最初に、フラッシュメモリ２２に記憶させたＥＰＧ情報の中から、タイトル文字列情報を取り出す（ステップＳ１）。

【００５１】

続いて、そのタイトル文字列情報が示す複数の番組のタイトルから、タイトル用キーワード辞書に登録されている文字列（‘ゴルフ’，‘サッカー’，‘温泉’，‘囲碁’，‘将棋’，‘映画’といったサブジャンルを示す文字列等）を探す。そして、それらの番組のタイトルのうち、このタイトル用キーワード辞書に登録されている文字列が含まれているタイトルの文字列全体を、キーワードの抽出対象とする（ステップＳ２）。

【００５２】

続いて、ステップＳ２でキーワードの抽出対象としたタイトルのうち、タイトル用除外文字列辞書に登録されている文字列（‘映画’，‘ＢＳ’等）の部分をスペースで置換する（ステップＳ３）。

【００５３】

続いて、ステップＳ３を経たタイトルの文字列から、図５に示すようなタイト

ル用の抽出ルールでキーワードを抽出する（ステップＳ４）。

【００５４】

このタイトル用抽出ルールでは、そのタイトルの文字列が平仮名，カタカナ，漢字，数字，アルファベット以外の特殊文字（スペースや×や「」等）で区切られていない場合は、そのタイトルの文字列をそのままキーワードとして抽出する。他方、そのタイトルの文字列がこうした特殊文字で区切られている場合は、特殊文字で区切られている各文字列のうちの２文字以上の文字列をそれぞれキーワードとして抽出する。

【００５５】

但し、‘・’（中点）は特殊文字としては扱わない。そして、キーワードとして抽出した文字列の先頭か末尾に‘・’（中点）が存在する場合には、‘・’（中点）を除いた部分をキーワードとする。

【００５６】

最後に、ステップＳ４で抽出したキーワードを、タイトル文字列情報中のキーワードのリストとしてフラッシュメモリ２２に記憶させる（ステップＳ５）。

【００５７】

次に、図４は、詳細文字列情報からキーワードを抽出する処理であり、最初に、フラッシュメモリ２２に記憶させたＥＰＧ情報の中から、詳細文字列情報を取り出す（ステップＳ１１）。

【００５８】

続いて、その詳細文字列情報から、詳細情報用キーワード辞書に登録されている文字列（有名な人の名称等）を探す。そして、その詳細文字列情報のうち、この詳細情報用キーワード辞書に登録されている文字列をキーワードとして抽出するとともに、その文字列の部分を半角スペースで置換する（ステップＳ１２）。

【００５９】

続いて、ステップＳ１２を経た詳細文字列情報の文字列のうち、詳細情報用除外文字列辞書に登録されている文字列（‘ゲスト’，‘以上’，‘監督’等）の部分を半角スペースで置換する（ステップＳ１３）。

【００６０】

続いて、ステップＳ１３を経た詳細文字列情報の文字列から、図６に示すような詳細文字列情報用の抽出ルールでキーワードを抽出する（ステップＳ１４）。

【００６１】

この詳細文字列情報用抽出ルールでは、基本的には、平仮名とカタカナと漢字と数字とアルファベットとその他の字種の文字とを互いに分離する字種切り法を利用する。

【００６２】

但し、カタカナとアルファベットとは同一の字種として扱う（分離しない）。また、‘・’（中点）は、その直前の文字がカタカナ，アルファベットである場合にはそれぞれカタカナ，アルファベットとして扱う（分離しない）。

【００６３】

そして、分離した各文字列のうち、平仮名のみの文字列，２文字以下の漢字のみの文字列，６文字以上の漢字のみの文字列を除いた文字列を、それぞれキーワードとして抽出する。但し、キーワードとして抽出した文字列の先頭か末尾に‘・’（中点）が存在する場合には、‘・’（中点）を除いた部分をキーワードとする。

【００６４】

最後に、ステップＳ１２で抽出したキーワードと、ステップＳ１４で抽出したキーワードを、詳細文字列情報中のキーワードのリストとしてフラッシュメモリ２２に記憶させる（ステップＳ１５）。

【００６５】

次に、この番組記録再生装置２において番組検索のためのキーワードが抽出される様子を、具体例を挙げて説明する。

【００６６】

ユーザーの選局操作時やユーザーの録画予約操作に基づく録画時に多重分離器１４から送られてフラッシュメモリ２２に記憶されたＥＰＧ情報中のタイトル文字列情報に、例えば次のようなタイトルが含まれていたとする（但し□□，△△はプロ野球チーム名である）。

愛のから騒ぎ

プロ野球中継　□□×△△

ＢＳ映画「スペース・ウォーズ」

【００６７】

すると、図３の処理では、‘愛’，‘プロ野球’，‘映画’という文字列がタイトル用キーワード辞書に登録されているので、ステップＳ２で、これらの各タイトルについて、それぞれそのタイトルの文字列全体がキーワードの抽出対象となる。

【００６８】

そして、これらのタイトルのうち、ＢＳ映画「スペース・ウォーズ」については、ステップＳ３で、‘ＢＳ’の部分と‘映画’の部分とがスペースで置換される。

【００６９】

また、これらのタイトルのうち、プロ野球中継　□□×△△については、‘プロ野球中継’と‘□□’との間にスペース（特殊記号）が存在し、‘□□’と‘△△’との間にも×（特殊記号）が存在するので、ステップＳ４で、文字列‘プロ野球中継’，‘□□’，‘△△’がそれぞれキーワードとして抽出される。

【００７０】

また、これらのタイトルのうち、‘ＢＳ’，‘映画’の部分をスペース置換された「スペース・ウォーズ」については、「」（特殊記号）で区切られており、また‘・’（中点）は特殊記号として扱わないので、ステップＳ４で、本来の映画のタイトルそのものである‘スペース・ウォーズ’がキーワードとして抽出される。

【００７１】

また、これらのタイトルのうち、愛のから騒ぎは、特殊記号で区切られていないので、ステップＳ４で、タイトルそのものである‘愛のから騒ぎ’がキーワードとして抽出される。

【００７２】

したがって、ステップＳ５では、以下の文字列が番組検索用のキーワードとしてフラッシュメモリ２２に記憶される（前述のように□□，△△はプロ野球チー

ム名である）。

愛のから騒ぎ

プロ野球中継

□□

△△

スペース・ウォーズ

【００７３】

このようにして、愛のから騒ぎ，スペース・ウォーズというように特殊文字で区切られていないタイトルについては、図３の処理により、そのタイトルに含まれる複数の文字列がばらばらのキーワードとして抽出されることなく、そのタイトルそのものがそのままの形でキーワードとして抽出される。

【００７４】

こうした特殊文字で区切られていないタイトルは、そのタイトルに含まれる‘愛’，‘スペース’等の個々の文字列は意味が広すぎて番組検索のためのキーワードとしてあまり役立たず（検索結果が非常に多くなり）、タイトルそのものとしてはじめて番組の効率的な検索のためのキーワードとして役立つことが多い。したがって、ユーザーは、抽出されたキーワード（タイトルそのもの）を用いて、番組を効率的に検索することができるようになる。

【００７５】

また、スペース・ウォーズという映画のタイトル文字列については、タイトル文字列情報においてこのタイトルに付加されていた‘ＢＳ’，‘映画’というような番組検索のためには一般的過ぎる文字列がキーワードに含まれていないとともに、このタイトル文字列情報においてこのタイトルを囲っていた「」もキーワードに含まれていない。したがって、ユーザーは番組を効率的に検索することができるようになる。

【００７６】

また一方では、プロ野球中継　□□×△△というように特殊文字（スペースや‘×’）で区切られているタイトルについては、図３の処理により、特殊文字で区切られている個々の文字列であるプロ野球中継，□□，△△がそれぞれキーワ

ードとして抽出される。

【００７７】

こうした特殊文字で区切られているタイトルは、その特殊文字で区切られている個々の文字列がそれぞれ番組検索のためのキーワードとして役立ち、タイトルそのものとしては限定されすぎて番組検索のためのキーワードとしてあまり役立たない（対戦チーム（□□や△△の具体名）が異なれば別のタイトルになってしまうので、検索結果がゼロまたは非常に少なくなる）ことが多い。したがって、ユーザーは、抽出されたキーワード（特殊文字で区切られている個々の文字列）を用いて、やはり番組を効率的に検索することができるようになる。

【００７８】

他方、図４の処理では、フラッシュメモリ２２に記憶されたＥＰＧ情報中のこれらのタイトルの番組の詳細文字列情報から、詳細情報用キーワード辞書に登録されている有名人（愛のから騒ぎという番組の司会者，ゲストや、映画スペース・ウォーズに出演している俳優）の名称等が、ステップＳ１２でキーワードとして抽出される。

【００７９】

その際、苗字が漢字で名前が平仮名やカタカナの有名人の名称（例えば石田あかりという名称）もこの詳細情報用キーワード辞書に登録されているので、そうした有名人の名称もキーワードとして抽出される。

【００８０】

また、インターネット経由でダウンロードした最新の詳細情報用キーワード辞書も用いるので、最近有名になったばかりの人の名称もキーワードとして抽出される。

【００８１】

また、その詳細文字列情報のうち、その有名人の名称等の部分と、詳細情報用除外文字列辞書に登録されている文字列（‘ゲスト’，‘以上’，‘監督’等）の部分とが、ステップＳ１２及びＳ１３で半角スペースに置換される。

【００８２】

そして、このスペース置換された詳細文字列情報の文字列から、ステップＳ１

４で、図６に示したルールによってキーワードが抽出される。

【００８３】

その際、カタカナとアルファベットとは同一の字種として扱われるとともに‘・’（中点）はその直前の文字がカタカナ，アルファベットである場合にはそれぞれカタカナ，アルファベットとして扱われるので、名前と苗字との間に‘・’（中点）が挿入された外国人名（例えばＢ・ドゥーリー）もキーワードとして抽出される。

【００８４】

また、最新の詳細情報用キーワード辞書にもまだ登録されていない人（例えばデビューしたばかりの無名の芸能人）の名称でも、平仮名のみの名称や２文字以下の漢字のみの名称や６文字以上の漢字のみの名称（すなわち人名としてあまりなさそうな名称）でなければキーワードとして抽出される。

【００８５】

また、‘ゲスト’，‘以上’，‘監督’といったような番組検索のためには不適切な文字列は、スペース置換されているのでキーワードとして抽出されることはない。

【００８６】

これにより、ステップＳ１５では、苗字が漢字で名前が平仮名やカタカナの有名人名や、最近有名になったばかりの人の名称や、名前がアルファベットで表記され苗字がカタカナで表記された外国人名や、名前と苗字との間に‘・’（中点）が挿入された外国人名も、番組検索用のキーワードとしてフラッシュメモリ２２に記憶される。したがって、ユーザーは、抽出されたキーワードを用いて、番組を効率的に検索することができるようになる。

【００８７】

なお、図３，図４の処理によってフラッシュメモリ２２に記憶させたキーワードをユーザーが番組検索のために用いる方法としては、例えば、リモコン２８による所定の操作に基づき、ＣＰＵ１９が、番組検索用画面（キーワードを一覧表示するとともにユーザーがその中の所望のキーワードを選択して検索を指示するための画面）の映像信号を作成して映像信号処理回路１７，映像出力端子２９を

経て表示装置３に送るといったような、適宜の方法をとればよい。

【００８８】

以上のように、この番組記録再生装置２では、ＥＰＧ情報中のタイトル文字列情報からのキーワードの抽出と詳細文字列情報からのキーワードの抽出とを、それぞれの情報に合せて互いに異なるキーワード辞書とルールとで行うことにより、小さなサイズのプログラムや辞書で精度よくキーワードを抽出することができるようになっている。

【００８９】

これにより、ＣＰＵ１９の処理能力やメモリ（ＲＯＭ２０やフラッシュメモリ２２等）の容量がさほど大きくなくても、ＥＰＧ情報中のタイトル文字列情報及び詳細文字列情報から、ユーザーが番組を検索するためのキーワードを効率よく且つ精度よく自動的に抽出することができるようになっている。

【００９０】

なお、以上の例では、デジタルテレビジョン放送の番組を記録・再生する装置に本発明を適用している。しかし、これに限らず、アナログテレビジョン放送の番組を記録・再生する番組記録再生装置にも本発明を適用してよいことはもちろんである。

【００９１】

図７は、本発明を適用したアナログテレビジョン放送用の番組記録再生装置のハードウェア構成を示すブロック図である。アンテナ３１で受信されて番組記録再生装置４１に入力したアナログ放送信号中の映像・音声信号は、チューナ４２で周波数帯を選択され、ＭＰＥＧエンコーダ４３で符号化される。

【００９２】

テレビ番組の視聴時には、この符号化された映像・音声データは、ＭＰＥＧデコーダ４７で復号されて、番組記録再生装置４１から表示装置６１に送られる。

【００９３】

他方、テレビ番組の記録時には、ＭＰＥＧエンコーダ４３で符号化された映像・音声データは、バス４４を介して主記憶装置４５に送られて、主記憶装置４５に記録される。

【００９４】

そして、再生時には、主記憶装置４５から読み出された映像・音声データが、バス４４を介してＭＰＥＧデコーダ４７に送られ、ＭＰＥＧデコーダ４７で復号されて、番組記録再生装置４１から表示装置６１に送られる。

【００９５】

また、チューナ４２で周波数帯を選択されたアナログ放送信号から、ＥＰＧ取得モジュール４６でＥＰＧ情報が取得される。このＥＰＧ情報も、バス４４を介して主記憶装置４５に送られて、主記憶装置４５に記憶される。

【００９６】

また、インターネット７１と接続するための通信インターフェース４８，ＲＯＭ４９，主記憶装置５０，補助記憶装置５１，ＭＰＥＧデコーダ４７が、互いにバス５２で結ばれている。

【００９７】

この番組記録再生装置４１でも、前述のようなタイトル用キーワード辞書，タイトル用除外文字列辞書，詳細情報用キーワード辞書及び詳細情報用除外文字列辞書がＲＯＭ４９に格納されている（詳細情報用キーワード辞書に関してはインターネット経由で専用のサイトから最新のものをダウンロードして補助記憶装置５１にも記憶させる）とともに、番組記録再生装置４１全体を制御するＣＰＵ５３が、図３，図４に示したのと同じキーワードの自動抽出処理をこれらの辞書及び主記憶装置４５内のＥＰＧ情報を用いて行い、抽出したキーワードを補助記憶装置５１に記憶させる。

【００９８】

この番組記録再生装置４１でも、図１，図２の番組記録再生装置２について説明したのと全く同様にして、ＥＰＧ情報中のタイトル文字列情報からのキーワードの抽出と詳細文字列情報からのキーワードの抽出とを、それぞれの情報に合せて互いに異なるキーワード辞書とルールとで行うことにより、小さなサイズのプログラムや辞書で精度よくキーワードを抽出することができる。

【００９９】

これにより、ＣＰＵ５３の処理能力やメモリ（ＲＯＭ４９や補助記憶装置５１

等）の容量がさほど大きくなくても、ＥＰＧ情報中のタイトル文字列情報及び詳細文字列情報から、ユーザーが番組を検索するためのキーワードを効率よく且つ精度よく自動的に抽出することができる。

【０１００】

また、以上の例では、表示装置とは別体となった番組記録再生装置に本発明を適用している。しかし、これに限らず、この番組記録再生装置と表示装置とが一体となったテレビジョン受信機や、番組の記録再生機能を有しないテレビジョン受信機にも本発明を適用してよい。

【０１０１】

また、以上の例では、ＥＰＧ情報中の番組のタイトル文字列情報，詳細文字列情報からのキーワードの検索のために本発明を適用している。しかし、これに限らず、テレビジョン番組以外のコンテンツ（例えばインターネット経由で配信されるコンテンツ）のタイトル文字列情報，詳細文字列情報からのキーワードの検索のためにも本発明を適用してよい。

【０１０２】

また、本発明は、以上の例に限らず、本発明の要旨を逸脱することなく、その他様々の構成をとりうることはもちろんである。

【０１０３】

【発明の効果】

以上のように、本発明によれば、ＣＰＵの処理能力やメモリの容量がさほど大きくない家電製品でも、ＥＰＧ情報のような番組のタイトル文字列情報及び詳細文字列情報から、ユーザーが番組を検索するためのキーワードを、効率よく且つ精度よく自動的に抽出することができるという効果が得られる。

【図面の簡単な説明】

【図１】

本発明を適用した番組記録再生装置を含むデジタルテレビジョン放送受信システムの概要を示す図である。

【図２】

図１の番組記録再生装置のハードウェア構成を示すブロック図である。

【図３】

図２のＣＰＵが実行するキーワードの自動抽出処理を示すフローチャートである。

【図４】

図２のＣＰＵが実行するキーワードの自動抽出処理を示すフローチャートである。

【図５】

図３の処理におけるキーワード抽出のためのルールを示す図である。

【図６】

図４の処理におけるキーワード抽出のためのルールを示す図である。

【図７】

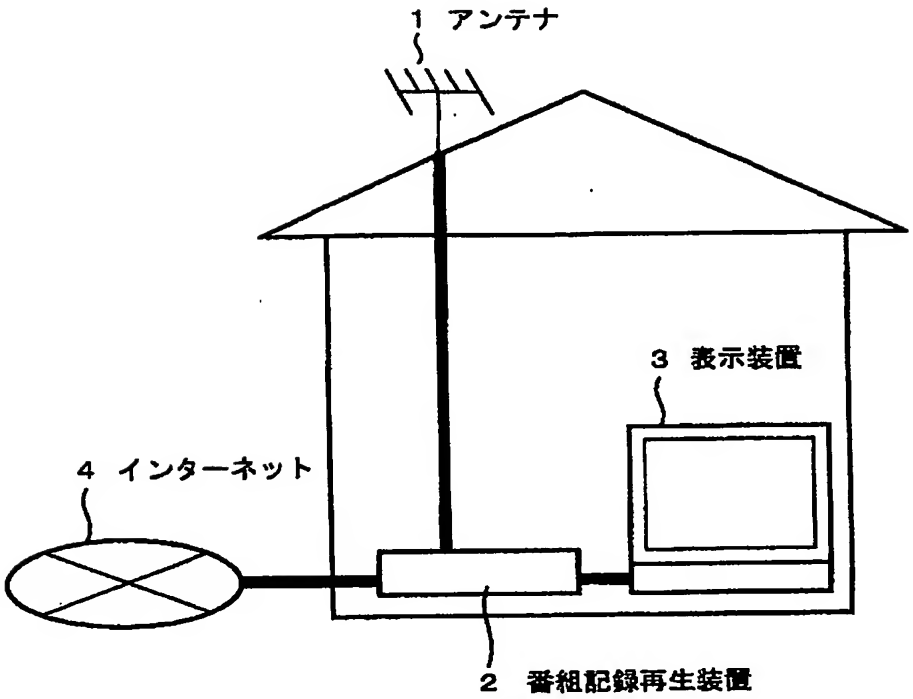
本発明を適用したアナログテレビジョン放送用の番組記録再生装置のハードウェア構成を示すブロック図である。

【符号の説明】

１，３１　アンテナ、　２，４１　番組記録再生装置、　３，６１　表示装置、　１１，４２　チューナ、　１２　復調器、　１３　デスクランブラ、　１４　多重分離器、　１５　映像デコーダ、　１６　音声デコーダ、　１７　映像信号処理回路、　１８　Ｄ／Ａ変換器１８、　１９　ＣＰＵ、　２０，４９　ＲＯＭ、　２１　ＲＡＭ、　２２　フラッシュメモリ、　２３　リモートコントローラ用のインターフェース、　２４　ＨＤＤ用のインターフェース、　２５，４８　インターネット接続用の通信インターフェース、　２６　システムバス、　２７　ＨＤＤ、　２８　リモコン、　４３　ＭＰＥＧエンコーダ、　４４，５２　バス、　４５，５０　主記憶装置、　４６　ＥＰＧ取得モジュール、　４７　ＭＰＥＧデコーダ、　５０　主記憶装置、　５１　補助記憶装置

【書類名】　　図面

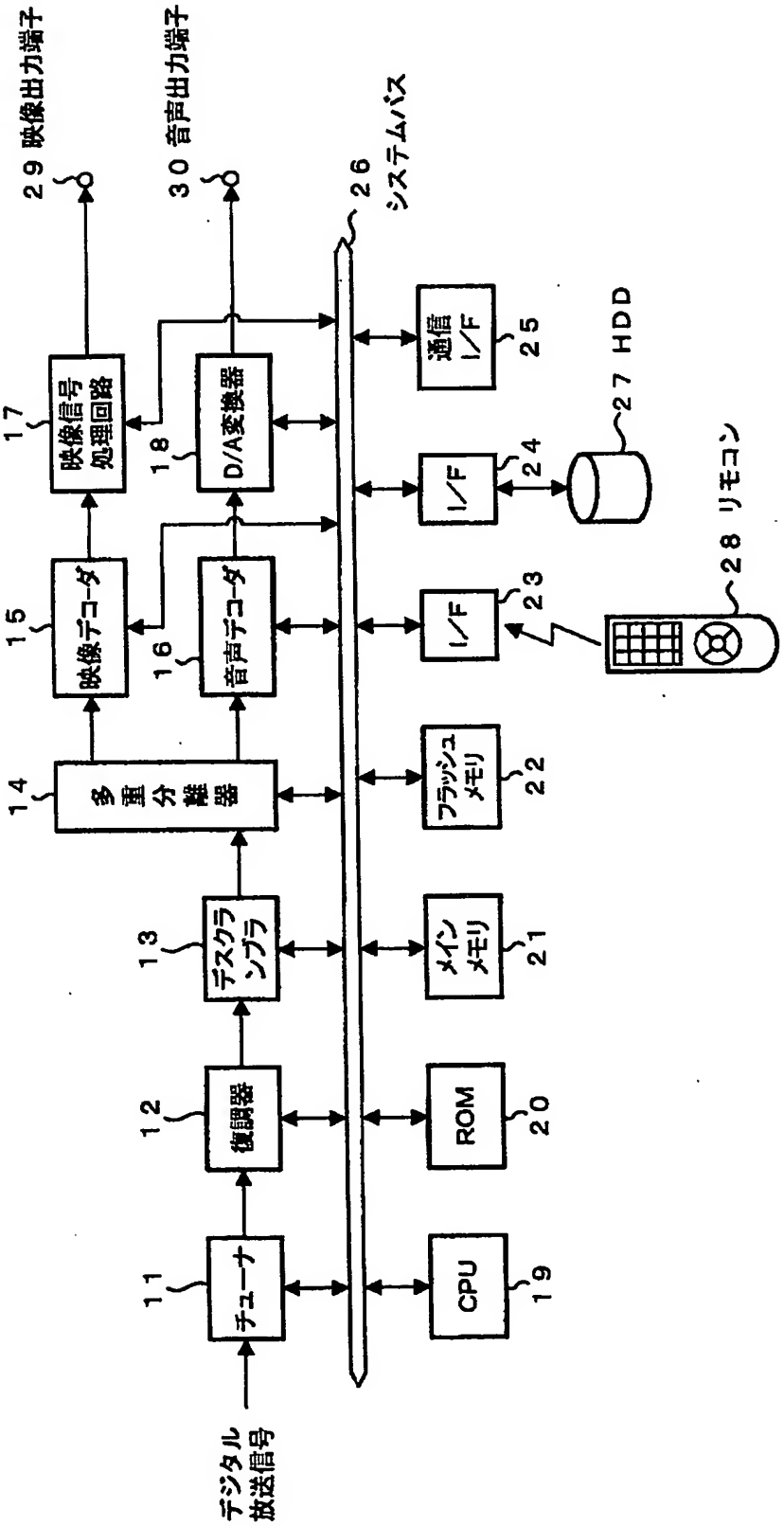
【図１】



本発明の番組記録再生装置を含む放送受信システム

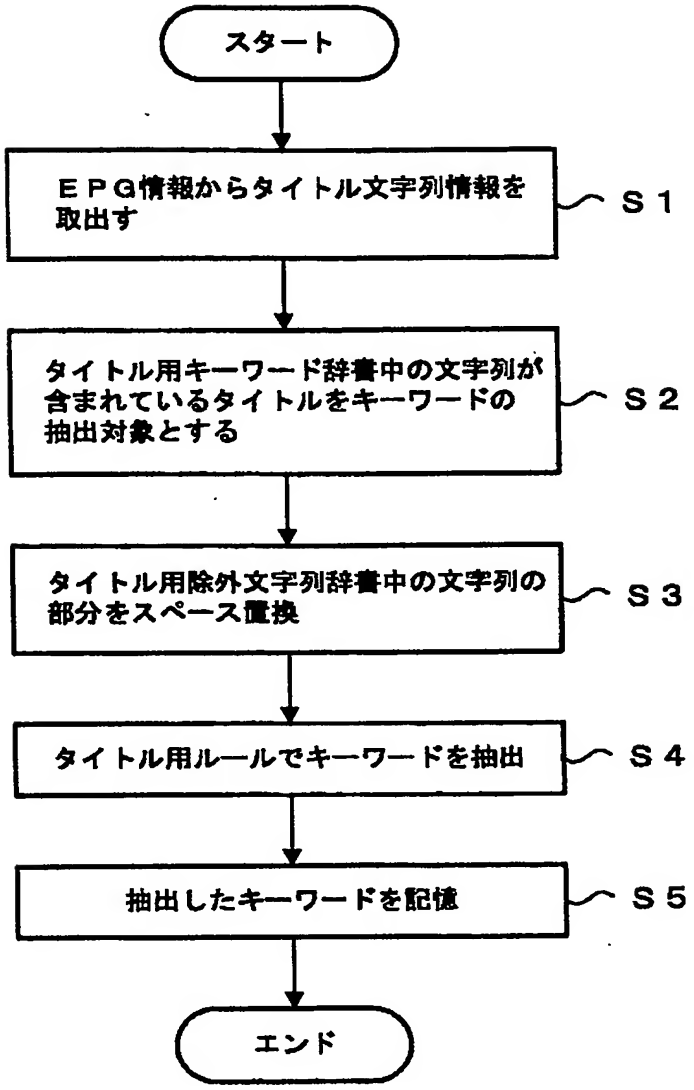
出証特２００３－３０４４５２０

【図２】



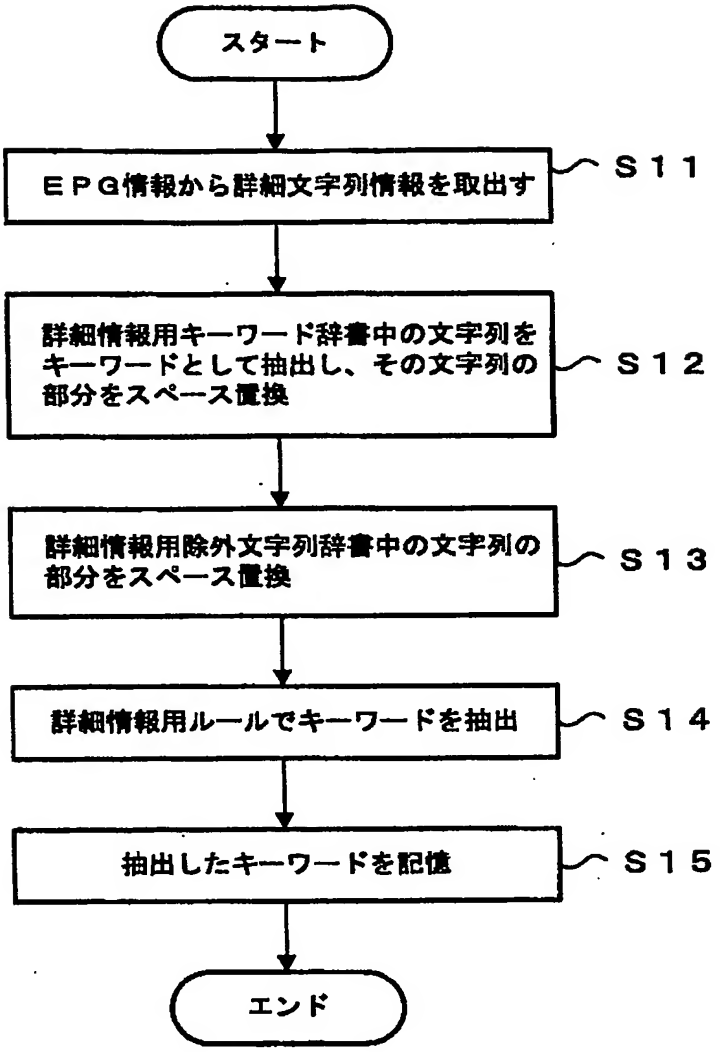
番組記録再生装置の構成

【図３】



キーワード自動抽出処理（タイトル文字列情報から）

【図４】



キーワード自動抽出処理（詳細文字列情報から）

【図５】

特殊文字で区切られていないタイトルの文字列をそのままキーワードとして抽出
特殊文字で区切られているタイトルの各文字列のうち、２文字以上の文字列をキーワードとして抽出
’・’（中点）は特殊文字として扱わない。抽出した文字列の先頭か末尾に、’・’が存在する場合には、’・’を除いた部分をキーワードとする

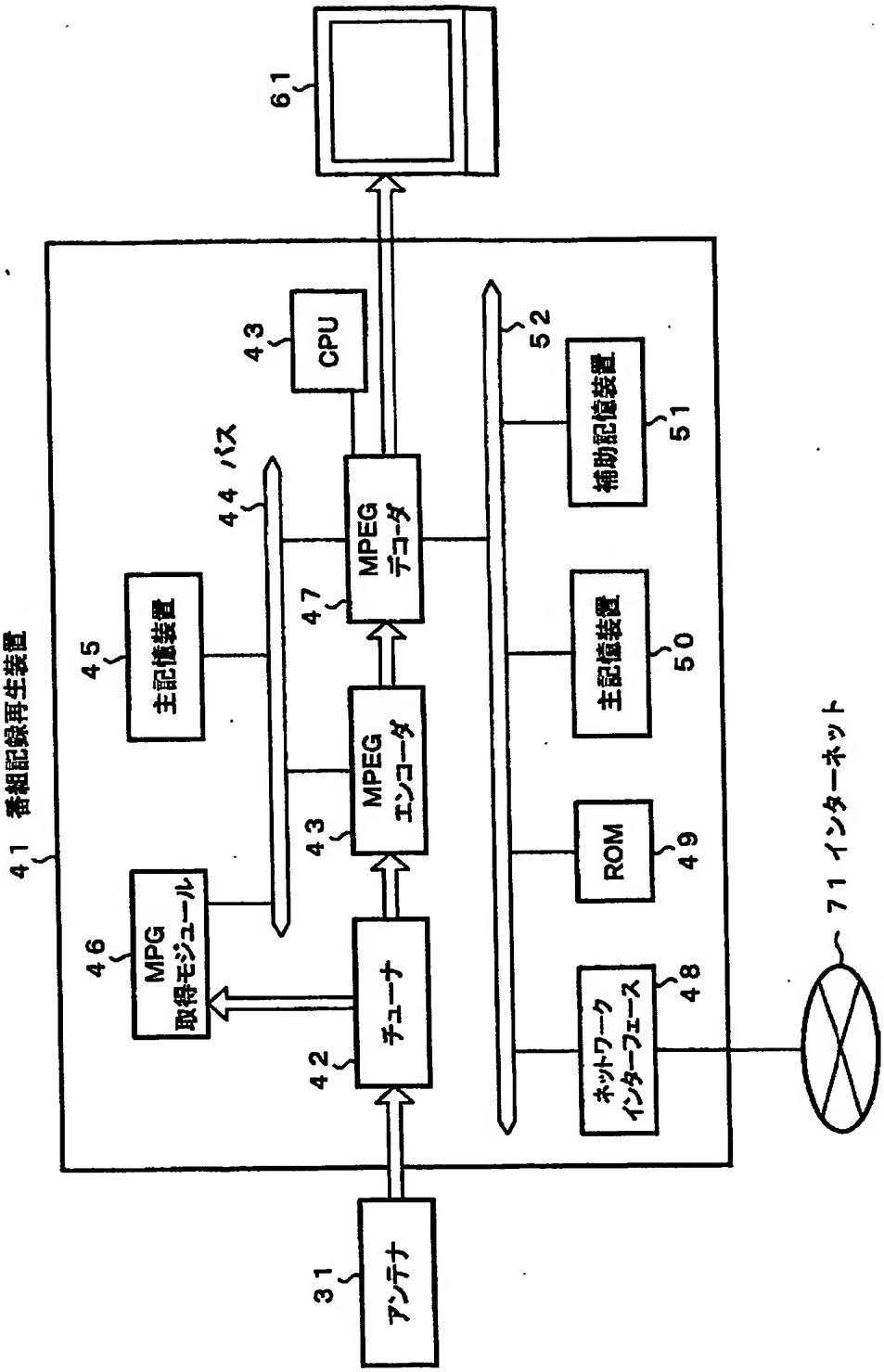
キーワード抽出ルール（タイトル用）

【図６】

字種切り法を利用
カタカナとアルファベットは同一字種として扱う
’・’は、直前の文字がカタカナ、アルファベットのとき、それぞれカタカナ、アルファベットとして扱う
分離した文字列のうち、平仮名のみ、２文字以下の漢字のみ、６文字以上の漢字のみを除いた文字列をキーワードとして抽出

キーワード抽出ルール（詳細文字列情報用）

【図７】


61
41 番組記録再生装置
43 CPU
52
51 補助記憶装置
44 バス
MPEG デコーダ
47
45 主記憶装置
50 主記憶装置
MPEG エンコーダ
43
ROM
49
46 MPG 取得モジュール
チューナ
42
ネットワーク インターフェース
48
71 インターネット
31 アンテナ

【書類名】　　要約書

【要約】

【課題】　ＣＰＵの処理能力やメモリの容量がさほど大きくない家電製品でも、ＥＰＧ情報のようなコンテンツのタイトル情報及び詳細情報から、ユーザーがコンテンツを検索するためのキーワードを、効率よく且つ精度よく自動的に抽出できるようにする。

【解決手段】　コンテンツのタイトル情報から、サブジャンルを示す文字列が登録されている第１のキーワード辞書を用いてキーワードの抽出を行う第１の抽出手段１９と、このコンテンツの詳細情報から、人名が登録されている第２のキーワード辞書を用いたキーワードの抽出と、字種切り法を利用したキーワードの抽出とを行う第２の抽出手段１９とを備える。

【選択図】　　図２

認定・付加情報

特許出願の番号	特願２００２－２２１６９８
受付番号	５０２０１１２５６１２
書類名	特許願
担当官	第一担当上席　００９０
作成日	平成１４年　７月３１日

＜認定情報・付加情報＞

【特許出願人】

【識別番号】	000002185
【住所又は居所】	東京都品川区北品川６丁目７番３５号
【氏名又は名称】	ソニー株式会社

【代理人】　申請人

【識別番号】	100122884
【住所又は居所】	東京都新宿区西新宿１丁目８番１号　新宿ビル　信友国際特許事務所
【氏名又は名称】	角田　芳末

【選任した代理人】

【識別番号】	100113516
【住所又は居所】	東京都新宿区西新宿１丁目８番１号　新宿ビル　松隈特許事務所
【氏名又は名称】	磯山　弘信

次頁無

出証特２００３－３０４４５２０

出願人履歴情報

識別番号　〔０００００２１８５〕

１．変更年月日　１９９０年　８月３０日
［変更理由］　新規登録
住　所　東京都品川区北品川６丁目７番３５号
氏　名　ソニー株式会社

２．変更年月日　２００３年　５月１５日
［変更理由］　名称変更
住　所　東京都品川区北品川６丁目７番３５号
氏　名　ソニー株式会社